19 place Marguerite
Perey 91120 Palaiseau

Extraction de texte dans les documents mixtes par approche deep-learning

IP-Paris/Telecom Paris
Proposition de stage 2020

Contexte

La recherche en analyse de documents et reconnaissance de l'écriture est un domaine très actif depuis une vingtaine d'années. Elle a fait émerger des modèles comme les modèles de Markov cachés (HMMs), et plus récemment des modèles à base de réseaux de neurones profonds. Ceux-ci très largement utilisés maintenant dans les domaines de la reconnaissance de la parole, la vision par ordinateur et pour toutes sortes d'applications. En ce qui concerne la reconnaissance d'écriture, les architectures profondes et récurrentes pour l'analyse de séquences sont maintenant très populaires (BLSTMs notamment).

Un système de reconnaissance d'images de documents, commence généralement par des prétraitements: nettoyage, réduction des déformations (dewarping), redressement. Puis on segmente le document en différentes couches. Cette segmentation fait intervenir des traitements qui cherchent à séparer la couche imprimée de la couche manuscrite [Belaid et al. 2013], repérer les tableaux, les images, les graphiques. La segmentation imprimé manuscrit reste encore un problème ouvert pour les documents mixtes, hétérogènes, voire multi-lingues. Cette segmentation permet de transmettre, suivant les cas, les images de texte à un classifieur spécifique au manuscrit ou un classifieur dédié à l'imprimé. A notre connaissance, il y a peu d'approches de type deep learning appliquées à ce problème [Chen et al. 2017][Huang et al. 2019] .

Les approches de type deep learning dans le domaine de la vision, notamment celles relatives à la détection de texte dans les scènes naturelles [Zhou et al. 2019] [Björklund et al. 2019] [Jaderberg et al. 2014], sont prometteuses pour l'analyse des images de documents. Leur mise en oeuvre nécessite de grandes bases de données annotées. La base Maurdor [Brunessaux et al. 2014], ainsi que la base IAMonDo [Indermühle et al. 2010] ont été construites pour permettre le développement et l'évaluation de systèmes de

19 place Marguerite
Perey 91120 Palaiseau

segmentation et/ou de reconnaissance. Ces bases pourront être utiles dans le cadre de ce stage.

L'objectif de ce stage est de développer de nouvelles approches de type deep learning pour la localisation des éléments textuels dans des documents mixtes réels. Nous nous intéresserons à des documents mixtes de type de constats d'accident et formulaires. Ces documents contiennent du texte imprimé en différentes tailles de police, des éléments manuscrits textuels et graphiques. Il s'agira de segmenter (localiser) ces éléments par une boîte englobante, en vue de leur reconnaissance par des classifieurs adaptés.

Comité d'encadrement

Laurence Likforman-Sulem, Telecom Paris/Institut Polytechnique de Paris http://www.telecom-paristech.fr/~lauli/

Attilio Fiandrotti, Telecom Paris/Institut Polytechnique de Paris.

Lieu du stage : Laboratoire LTCI, département IDS

19 place Marguerite Perey

91120 Palaiseau

Profil recherché

Stage de Master ou de fin d'études. Compétences en Machine learning, Deep learning, programmation (Python).

Références

A. Belaïd, Santosh K.C., V. Poulain d'Andecy. Handwritten and Printed Text Separation in Real Document. IAPR International Conference on Machine Vision Applications, 2013, Kyoto, Japan.

T Björklund, A Fiandrotti, M Annarumma, G Francini, E Magli, Robust license plate recognition using neural networks trained on synthetic images, Pattern Recognition 93, 134-146, 2019.

S. Brunessaux et al. The Maurdor Project: Improving Automatic Processing of Digital Documents, IAPR International Workshop on Document Analysis Systems, 2014.

K. Chen, M. Seuret, J. Hennebert, R. Ingold, Convolutional Neural Networks for Page Segmentation of Historical Document Images, ICDAR 2017.

19 place Marguerite
Perey 91120 Palaiseau

E. Granell, E. Chammas, L. L. Likforman-Sulem, C.-D. Martinez-Hinarejos, C. Mokbel, B.-I Cirstea, Transcription of Spanish Historical Handwritten Documents with Deep Neural Network, Journal of Imaging special issue on Document Image Processing.

Yaoxiong Huang, Zecheng Xie, Lianwen Jin, Yuanzhi Zhu and Shuaitao Zhang, Adversarial Feature Enhancing Network for End-to-End Handwritten Paragraph Recognition, ICDAR 2019.

E. Indermühle, M. Liwicki, and H. Bunke: IAMonDo-database: an Online Handwritten Document Database with Non-uniform Contents. Proc 9th Int. Workshop on Document Analysis Systems, 2010.

M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman. "Synthetic data and artificial neural networks for natural scene text recognition." arXiv preprint arXiv:1406.2227, 2014.

Xiangyu Zhou et al. Deep Residual Text Detection Network for Scene Text, ICDAR 2017.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, EAST: An Efficient and Accurate Scene Text Detector, ArXiv:1704.03155v2 [cs.CV] 10 Jul 2017